Recovering Transitions From Repeated Cross-Sectional Samples

Eisinga, Rob

Published in:
Methodology

DOI:
10.1027/1614-2241.4.4.139

Recovering Transitions From Repeated Cross-Sectional Samples

Rob Eisinga

Radboud University Nijmegen, The Netherlands

Abstract. This paper proposes a dynamic Markov model for the estimation of binary state-to-state transition probabilities from a sequence of independent cross-sectional samples. It discusses parameter estimation and inference using maximum likelihood (ML) methodology. The model is illustrated by the application of a three-wave panel study on pupils' interest in learning physics. These data encompass more information than what is used to estimate the model, but this surplus information allows us to assess the accuracy and the precision of the transition estimates. Bootstrap and Bayesian simulations are used to evaluate the accuracy and the precision of the ML estimates. To mimic genuine cross-sectional data, samples of independent observations randomly drawn from the panel are also analyzed.

Keywords: repeated cross-sections, Markov transition model, 2×2 table, ecological inference

The purpose of this paper is to present a nonstationary, heterogeneous Markov model that affords a transitional analysis of survey data from a time series of independent cross-sectional samples. This model is most suitable for examining individual-level binary transitions between states over time, under first-order Markov assumptions, and the effects of time-constant and time-varying covariates on the transition intensities. We describe maximum likelihood (ML) parameter estimation and illustrate the model with an educational case study.

In its simplest formulation, the problem studied in this paper is to estimate the cell probabilities in a series of 2×2 tables, when only marginal counts are observed and the entries are unavailable. This problem has attracted the attention of researchers from various disciplines and the aspects of the problem have recently become of increasing interest in several literatures with varying concerns and terminology. Many statisticians have considered the amount of information that the marginal totals provide about the interior cells and have derived ML estimates of the cell probabilities under different sampling schemes (e.g., Haber, 1989; Hamdan & Nasro, 1986; Kocherlakota & Kocherlakota, 1992; McCullagh & Nelder, 1992; Plackett, 1977). There has also been an upsurge of interest in the past several years on the problem of making individual-level inference from aggregate-level data or, more formally, on point identification of $P(y|x, z)$ from information on only $P(y|x)$ and $P(z|x)$. Various inferential methods to aid in the solution of such ecological inference problems have been proposed in many disciplines, including political science (Achen & Shively, 1995; King, 1997; King, Rosen, & Tanner, 2003), epidemiology (Richardson & Montfort, 2000), statistics (Wakefield, 2003), marketing (Böckenholt & Dillon, 2000), and econometrics (Cross & Manski, 2002). See the special issue on ecological analysis in the *Journal of the Royal Statistical Society, Series A* (2001, volume 164, issue 1) for other application areas. Further, there is a fair amount of recent activity in the development of confidentiality protection and statistical data disclosure limitation methods, and a central issue studied in this area is the potential to infer microdata from the sets of interlinked aggregate data (Dobra, Tebaldi, & West, 2003; Fienberg, 1997; Tebaldi & West, 1998). An important role in both disclosure limitation and ecological inference procedures is played by the lower and upper (Fréchet) bounds on cell counts in marginally constrained cross-classifications. Finally, considerable attention has recently been given to the use of maximum entropy procedures to recover the entries from incomplete contingency table data (e.g., Golan, Judge, & Miller, 1996; Golan, Judge, & Robinson, 1994; Judge, Miller, & Tam Cho, 2003). The key feature of these information theoretic approaches is to convert an ill-posed (underdetermined) inverse problem into a well-posed one and to use (Shannon's) entropy as a decision criterion for estimation and inference. Here, we restrict the attention to finding estimates by ML. Maximum entropy formulations for recovering the unknown transitions unobserved and unobservable in repeated cross-sections are addressed in a separate study.

In this paper we develop, apply, and verify a dynamic Markov model that can be used to recover binary-state transition probabilities at the individual level from repeated cross-sectional (RCS) sample survey data that lack direct information on individual turnover. The estimated transitions are allowed to vary over time and from subject to subject. The model may circumvent the need for individual-level panel data and their associated problems (e.g., nonrandom attrition), and overcome some deficiencies in aggregate time series (e.g., aggregation bias), by exploiting individual-level information from independent samples of cross-sections of the population collected over time. Also, a number of the proposed approaches assume tables with large marginal totals or population data, so that knowledge of $P(y|x)$ and $P(z|x)$ implies a deterministic bound on $P(y|x, z)$. We, in contrast, assume a sample situation and work with a series of relatively small-sized samples, wherein the observed data

DOI: 10.1027/1614-2241.4.4.139

are estimates of the true proportions. Throughout this paper we shall restrict ourselves to a two-state transition model, but this is not essential. Multi-state models can be developed analogously, at the cost of some extra complexity in the computations.

The remainder of this paper is organized as follows: Section "Interior Cells and Likelihood" briefly considers the wider statistical literature concerned with making inferences about interior cell probabilities from marginal totals. Section "Estimating Transitions From Repeated Cross-Section" outlines our proposed two-state Markov model for RCS data and discusses ML estimation and inference. Section "Application" applies the model to a published panel data set on pupils' interest in learning physics and verifies the results against the panel observations. Concluding remarks are offered in the "Discussion" section.

Interior Cells and Likelihood

Sampling Design and Likelihood Specification

We first consider the case in which the interior cells are observed. Assume that we are sampling (with replacement) from a multinomial population and that the observed sample is available in the form of a 2×2 contingency table with completely classified observations: that is, all four cell counts given and no margins fixed. The entries, y_{ij}, in each cell denote the number of observations in row i and column j, with $i, j = 0, 1$, and p_{ij}, the unknown parameters.

	0	1	
0	y_{00}	y_{01}	$n - y_{t-1}$
1	y_{10}	y_{11}	$y_{t-1} = y_{1.}$
	$n - y_t$	$y_t = y_{.1}$	n

	0	1	
0	p_{00}	p_{01}	$1 - p_{t-1}$
1	p_{10}	p_{11}	$p_{t-1} = p_{1.}$
	$1 - p_t$	$p_t = p_{.1}$	1

The probability distribution of y_{ij} obviously depends on the sampling scheme that was used to generate the table. With zero fixed margins and entries (Y_{00}, Y_{01}, Y_{10}, and Y_{11}) as random variables, the exact model is the four-cell multinomial distribution with parameters (p_{00}, p_{01}, p_{10}, and p_{11}). If the Y_{ij} take non-negative integer values and sum to (fixed) n, the probability is

$$P(Y_{00} = y_{00}, \ldots, Y_{11} = y_{11}) = \frac{n!}{\prod_{i,j} y_{ij}!} \prod_{i,j} p_{ij}^{y_{ij}}, \tag{1}$$

where $p_{ij} \in [0, 1]$ and $\sum_{i,j} p_{ij} = 1$. Let y_{t-1} and y_t denote the sums of the multinomial variates $y_{10} + y_{11}$ and $y_{01} + y_{11}$, respectively, and define the corresponding marginal probabilities by $p_{t-1} = p_{10} + p_{11}$ and $p_t = p_{01} + p_{11}$. The marginal distributions of the sums are binomial, here $Y_{t-1} \sim B(n, p_{t-1})$ and $Y_t \sim B(n, p_t)$. This binomial distribution also appears as the conditional distribution of a subset of the components of the multinomial random variable, given the values of the remaining components. Using this result, we have

$$\begin{aligned} P(Y_{00} = y_{00}, \ldots, Y_{11} = y_{11}) &= \binom{n}{y_{t-1}} p_{t-1}^{y_{t-1}} (1 - p_{t-1})^{n - y_{t-1}} \\ &\times \left\{ \binom{n - y_{t-1}}{y_{01}} \left(\frac{p_{01}}{1 - p_{t-1}} \right)^{y_{01}} \left(\frac{p_{00}}{1 - p_{t-1}} \right)^{y_{00}} \right. \\ &\times \left. \binom{y_{t-1}}{y_{11}} \left(\frac{p_{11}}{p_{t-1}} \right)^{y_{11}} \left(\frac{p_{10}}{p_{t-1}} \right)^{y_{10}} \right\}. \end{aligned} \tag{2}$$

Hence, the multinomial likelihood function for the 2×2 table factorizes into a marginal binomial random variable for the row totals and two conditional binomials for the two rows (Bishop, Fienberg, & Holland, 1975).

Now consider the situation where the two marginal totals are observed, but the individual cell counts are unavailable for some reason (e.g., sampling design and data security). Then three types of sampling, depending on the number of margins fixed by the sampling scheme (0, 1, or 2), give rise to the observed table (Barnard, 1947). The case of two fixed margins is relatively rare in applied statistics, but the other two occur frequently.

If the data consist of only marginal information and none of the margins is fixed, the bivariate binomial distribution may be employed to investigate two-dimensional random variables, where each of the marginal probability functions is binomial. If we write $g = y_{01}$ to indicate that the cell counts are unobserved, the joint probability mass function of Y_{t-1} and Y_t is given by

$$\begin{aligned} &P(Y_{t-1} = y_{t-1}, Y_t = y_t) \\ &= \sum_{g=u_0}^{u_1} \binom{n}{g, n - y_{t-1} - g, y_t - g, y_{t-1} - y_t + g} \\ &\times p_{01}^{g} (1 - p_{t-1} - p_{01})^{n - y_{t-1} - g} (p_t - p_{01})^{y_t - g} \\ &\times (p_{t-1} - p_t + p_{01})^{y_{t-1} - y_t + g}, \end{aligned} \tag{3}$$

where $u_0 = \max(0, y_t - y_{t-1})$ and $u_1 = \min(n - y_{t-1}, y_t)$. Given the data for pairs of observations (y_{t-1}, y_t) from a series of independent samples of size n (i.e., multiple tables), the parameters p_{t-1}, p_t, and p_{01} may be estimated by ML. Hamdan and Nasro (1986) and subsequently Kocherlakota and Kocherlakota (1992) studied this distribution in connection with 2×2 tables obtained by classifying each of a sample of n objects according to the presence or the absence of a pair of characteristics. They also provide estimation details such as derivatives of the likelihood with respect to the parameters.

Alternatively, with one margin fixed, inference can be accomplished by restricting the parameter space to the set of all 2×2 contingency tables that have the same fixed marginal sums as the observed table. Conditioning on the observed margins of the table affords a convenient way to eliminate nuisance parameters from the likelihood. Let $\mu = p_{01}/(1 - p_{t-1})$ and $\kappa = p_{11}/p_{t-1}$. In a comparative study (e.g., prospective trial) with two independent samples, the simplest model takes $Y_{01} \sim B(n - y_{t-1}, \mu)$ and $Y_{11} \sim B(y_{t-1}, \kappa)$ as the independent random variables. The probability of observing Y_t is then the convolution (e.g., Woodward & Palmer, 1997) of two independent binomials:

$$P(Y_t = y_t) = \sum_g \binom{n - y_{t-1}}{g} \mu^g (1-\mu)^{n-y_{t-1}-g} \times \binom{y_{t-1}}{y_t - g} \kappa^{y_t - g} (1-\kappa)^{y_{t-1}-y_t+g}, \quad (4)$$

where the summation is again over all g satisfying $\max(0, y_t - y_{t-1}) \leq g \leq \min(n - y_{t-1}, y_t)$. This double binomial has, in the context of 2×2 tables with unknown cell counts, been discussed by Eisinga (2008), Haber (1989), McCullagh and Nelder (1992), and Plackett (1977). McCue (1995) and Wakefield (2003) explicitly considered working with the convolution likelihood in the context of ecological inference analysis and presented normal approximations for tables with large marginal counts.

We finally note that on some occasions we may observe a 2×2 table where the row and column margins are both fixed by design (Ronald Fisher's hypothetical tea tasting experiment is the classic example). The sampling distribution of Y_{01} given the marginal totals is then an extended (or noncentral) hypergeometric (see, Agresti, 1992; Fisher, 1935; McCullagh & Nelder, 1992; Wakefield, 2003). Suppose Y_{01} and Y_{11} are independent binomial random variables with distributions $B(n - y_{t-1}, \mu)$ and $B(y_{t-1}, \kappa)$, respectively, and $\psi = \kappa(1-\mu)/\mu(1-\kappa)$ is the odds ratio, characterizing row-column nonindependence. Then the conditional distribution of Y_{01} given $Y_t = y_t$ and $Y_{t-1} = y_{t-1}$ is an extended hypergeometric with target parameter ψ, given by

$$\begin{aligned} P(Y_{01} = y_{01}) &= \frac{\psi^{y_{01}}}{\prod_{i,j} y_{ij}!} \left(\sum_{y_{01}} \frac{\psi^{y_{01}}}{\prod_{i,j} y_{ij}!} \right)^{-1} \\ &= \psi^{y_{01}} \binom{n - y_{t-1}}{y_{01}} \binom{y_{t-1}}{y_t - y_{01}} \\ &\quad \times \left\{ \sum_{g=u_0}^{u_1} \psi^g \binom{n - y_{t-1}}{g} \binom{y_{t-1}}{y_t - g} \right\}^{-1} \\ &\quad y_{01} = u_0, \ldots, u_1, \end{aligned} \quad (5)$$

where the summation index g ranges from $u_0 = \max(0, y_t - y_{t-1})$ to $u_1 = \min(n - y_{t-1}, y_t)$, that is, the possible values for y_{01} gives the marginal totals. Note that the distribution depends on the parameter ψ only.

An important factor for inference is whether it is admissible, in the sense of losing no information, to specify the sampling process only partially by treating part of the data as if they were not random. If some of the margins were not naturally fixed when the data were gathered, it may be inappropriate to fix them for the purpose of inference. We also note that no nondegenerate ML estimates exists for the above likelihoods if information is available from a single sample of size n (i.e., one table only). More specifically, for one observation on two marginal totals, the ML estimates of the cell probabilities always involve at least one cell whose estimated probability is equal to zero. ML estimation of the parameters thus requires information from multiple independent samples (see, Haber, 1989; Hamdan & Nasro, 1986; Kocherlakota & Kocherlakota, 1992; Plackett, 1977).

Repeated Cross-Sectional Design and Likelihood Function

In an RCS survey, the samples are drawn anew at every sampling occasion, making it impossible to observe lagged values of y_{it} for any individual over time. For the present discussion this implies that only one transition table margin is recorded in RCS data, the columns totals say, that the other margin is unrecorded, as are the interior cells. Write $f = y_{t-1}$ to indicate that the row sums are unavailable in repeated independent surveys. The joint probability function of Y_{t-1} and Y_t is then simply

$$P(Y_{t-1} = f, Y_t = y_t) = P(Y_{t-1} = f) \times P(Y_t = y_t | Y_{t-1} = f), \quad (6)$$

where $P(Y_t = y_t | Y_{t-1} = f)$ is the convolution distribution of Y_{01} and Y_{11} given by equation (4). If we conjecture that Y_{t-1} follows a binomial distribution, that is, $Y_{t-1} \sim B(n, p_{t-1})$, the marginal distribution of Y_t becomes

$$\begin{aligned} P(Y_t = y_t) = \sum_{f=0}^{n} \Bigg[\binom{n}{f} p_{t-1}^f (1-p_{t-1})^{n-f} \\ \times \sum_g \left\{ \binom{n-f}{g} \mu^g (1-\mu)^{n-f-g} \right. \\ \left. \times \binom{f}{y_t - g} \kappa^{y_t - g} (1-\kappa)^{f - y_t + g} \right\} \Bigg], \end{aligned} \quad (7)$$

and this corresponds to a binomial mass function. In other words, if Y_{01}, Y_{11}, and Y_{t-1} are all binomially distributed, the marginal distribution of Y_t is also binomial, with index parameter n and success probability p_t.

It is informative to consider the conditional distribution of Y_t given $Y_{t-1} = f$. The probability generating function (p.g.f.) of the distribution is given by

$$\Pi_{Y_t}(t|f) = \{(1-\mu) + \mu t\}^{n-f} \{(1-\kappa) + \kappa_t\}^f, \quad (8)$$

and this is recognized to be the p.g.f. of the convolution of $B(n-f, \mu)$ and $B(f, \kappa)$. Using this conditional distribution, the regression of Y_t given $Y_{t-1} = f$ is

$$E(Y_t|f) = \mu(n-f) + \kappa f = \mu n + (\kappa - \mu) f, \quad (9)$$

which is linear in f with regression coefficient $\kappa - \mu$. For this regression, the conditional variance can be determined as

$$\mathrm{Var}(Y_t|f) = \mu(1-\mu)n + \{\kappa(1-\kappa) - \mu(1-\mu)\} f, \quad (10)$$

which is also linear in f.

Further, lower and upper (Fréchet) bounds can be derived that define a feasible region within which μ and κ must lie. From equation (9), we have

$$\mu = \frac{E(Y_t|f)}{(n-f)} - \frac{f}{(n-f)} \kappa \quad \text{and} \quad \kappa = \frac{E(Y_t|f)}{f} - \frac{(n-f)}{f} \mu. \quad (11)$$

Since the left-hand sides of these identities are bounded on the [0, 1] interval, we have $\mu \in (L_\mu, U_\mu)$ and $\kappa \in (L_\kappa, U_\kappa)$, where the lower ($L$) and upper ($U$) bounds are defined by the min and max operators:

$$L_\mu = \max\left\{0, \frac{y_t - f}{n - f}\right\} \leq \mu \leq \min\left\{1, \frac{y_t}{n - f}\right\} = U_\mu \tag{12}$$

and

$$L_\kappa = \max\left\{0, \frac{y_t - (n - f)}{f}\right\} \leq \kappa \leq \min\left\{1, \frac{y_t}{f}\right\} = U_\kappa \tag{13}$$

(see, Chambers & Steel, 2001; King, 1997). At least two issues arise in computing lower and upper bounds from RCS data. One is that f is not observed. However, it is often not unreasonable to assume that the unknown y_{t-1} margin at time point t is equal to the observed margin of the previous sample at $t - 1$. To implement this assumption we may impute at time point t the marginal frequencies (or proportions) observed at $t - 1$. The other issue is that the lower and upper bounds provide deterministic information about μ and κ, conditional on the marginal data being correct. In cross-section samples, however, the y_t and imputed y_{t-1} are random observations. The lower and upper limits are, therefore, stochastic rather than deterministic bounds. The bounds obtained with RCS data may nevertheless provide useful statistical information. They may be used, for example, to get a preliminary impression of the range of estimates of the transition intensities. Taken the observed and imputed margins as given, we sometimes obtain sharp (albeit nondeterministic) bounds for μ and κ. Because the bounds are not known with certainty, it is informative to construct confidence intervals for the lower and upper limits, as will be done in the example application given below. Finally, it is important to note that our estimation procedure makes no use of equations (12) and (13). However, it does implicitly take into account the bounds and thereby restrict the ranges of feasible estimates of μ and κ, simply by constraining the estimated probabilities to lie within the unit interval.

Estimating Transitions From Repeated Cross-Section

Dynamic Transition Model

The data considered so far were given in the form of aggregate 2×2 tables. The problem of estimating transitions rates from either aggregate frequency data or aggregate proportions has been discussed extensively in several disciplines, particularly in statistics (e.g., Fingleton, 1997; Hawkins, Han, & Eisenfeld, 1996; Kalbfleish & Lawless, 1984, 1985; Kelton & Kelton, 1984; Lawless & McLeish, 1984; Lee, Judge, & Zellner, 1970; Li & Kwok, 1990), economics, and econometrics (e.g., Abowd & Zellner, 1985; Karantininis, 2002; Kelton, 1981; MacRae, 1977; Zepeda, 1995). However, an important feature of the model discussed here is that unlike these approaches, grouping of either the population or the cross-sectional sample data into aggregates (e.g., cohorts or other grouped data) need not be done. In fact, the variation in the microdata is utilized as part of the estimation procedure. For additional discussion of the similarities and dissimilarities between the traditional Markov chain aggregate data models and the current model, see Pelzer, Eisinga, and Franses (2001, 2002).

The Markov transition model that we outline in this section is designed to deal with individual-level observations. However, we first present the model in terms of aggregated data and subsequently explain the individual-level equivalent. We assume that the observed aggregate data consist of T independent cross-sections drawn at evenly spaced points in time with each being a random sample of n_t different individual units of some underlying population. This population is considered to be closed with respect to in- and out-migration and we assume that there are no births or deaths.

Let f again denote the unknown row count y_{t-1} of the cross-sectional survey observed at time point t. We consider f to be randomly drawn from $B(n_t, p_{t-1})$. The unknown cell counts Y_{01} and Y_{11} of cross-section t are also assumed to be binomially distributed: that is, $Y_{01,t} \sim B(n_t - f, \mu_t)$ and $Y_{11,t} \sim B(f, \kappa_t)$. For the resulting marginal distribution at time point t it follows that $Y_t \sim B(n_t, p_t)$. Note that the binomial parameters p_{t-1}, p_t, μ_t, and κ_t are probabilities that apply to all individual units of cross-section t. The parameters p_{t-1} and p_t denote the probability to be in state 1 at $t - 1$ and t, respectively, μ_t denotes the transition probability to enter state 1 at t given state 0 at $t - 1$, and κ_t the probability to stay in state 1 at t given state 1 at $t - 1$. The expectation of Y_t is again given by $E(Y_t|f) = n_t p_t = \mu_t(n_t - f) + \kappa_t f$ and dividing this expression by n_t leads to the following relationship between state probabilities at two consecutive points in time:

$$p_t = \mu_t(1 - p_{t-1}) + \kappa_t p_{t-1}. \tag{14}$$

Because the 0–1 Markov states are mutually exclusive and exhaustive and the row probabilities add to 1 in every time period, we may recursively apply equation (14) to get the following standard model:

$$p_t = \mu_t + \sum_{\tau=1}^{t-1}\left\{\mu_\tau \prod_{s=\tau+1}^{t}(\kappa_s - \mu_s)\right\} + p_0 \prod_{\tau=1}^{t}(\kappa_\tau - \mu_\tau). \tag{15}$$

Our aim is to establish consistent estimation of the unobserved interior cell counts or, equivalently, to estimate the underlying transition probabilities, given that the lagged dependent variable is not observed. This goal poses two challenges: (i) to overcome the identifiability problem inherent in equation (15) by imposing identifying constraints on the unobserved μ_t and κ_t transitions and (ii) to develop an estimation technique that efficiently uses the information we have and provides a basis for inference.

If equation (15) is estimated with independent data samples, the parameters are clearly not identifiable.

For a total of T cross-sections, equation (15) has $(T \times \mu_t + T \times \kappa_t + p_0 =) 2T + 1$ unknown parameters and T data points. Since the number of unknowns outnumber the number of knowns, the problem is underdetermined (or ill-posed). An obvious approach to overcome this ill-posed problem is re-parameterization, using less number of parameters. For example, the restriction $\mu_t = \mu$ and $\kappa_t = \kappa$ for all t reduces the number of unknowns to three and produces estimates for p_0, μ, and κ within the feasible range [0, 1].

Now suppose that instead of a single observed marginal count, the data consist of multiple y_{it} counts for each transition from $t-1$ to t and that each y_{it} pertains to a different group of n_{it} individual units having a unique set of values on a series of L covariates. The parameters in equations (14) and (15) would then have to be indexed by i. In the most extreme case of $n_{it} = 1$ for all i and t, each individual unit i has a unique set of covariate values and each observed y_{it} count is either 0 or 1, representing the observed value of Y_{it} for unit i. Equation (15) then becomes an individual-level Markov transition model. Further, with multiple y_{it} counts (or values) observed for each transition, the problem of estimating equation (15) becomes one of solving for $2t + 1$ unknown parameters for each observed y_{it} count. Because this problem is ill-posed, restraints have to be enforced on the temporal distribution of the transitions, cross-sectional distribution of the transitions or on both, for example, the constraint $\mu_{it} = \mu$ and $\kappa_{it} = \kappa$ for all i and t. A serious drawback of such homogeneous models is that they are based on the (often untested) assumption that subjects make transitions between states with common transition probabilities. When transition probabilities differ among sampling occasions of the population, among subsets of the population or among both, failure to account for the differences can result in biased estimates. In view of this, we adopt a remedy previously considered by Moffitt (1990, 1993). He suggests to replace the unobserved lagged dependent variable y_{it-1} by a predicted count $n_{it}p_{it-1}$, with p_{it-1} obtained using the values of the covariates at time period $t-1$. The central idea behind this approach is that it is often possible with RCS data to reconstruct past values of the covariates $\mathbf{x}_{it}$ by going back in time, starting from the cross-section at which they were observed. If so, the backcasted values may be used to predict p_{it-1}. This requires that the unit-specific covariates are either time-invariant (e.g., date of birth, gender, race, and completed education) or that lagged values of time-varying covariates can be obtained by retropolation (e.g., age and number of children of different ages). Macro-level aggregates may be used as well. If the repeated surveys contain such data, and the current and lagged $\mathbf{x}_{it}$ effectively affect μ_{it} and κ_{it}, the covariates can be employed to obtain current and backward predictions of the transition probabilities and hence the marginal probabilities.

While, in principle, μ_{it} and κ_{it} could be based on different sets of time-invariant or time-varying covariates, it is notationally most convenient that the same variables affect both transition processes, but with different parameters. With this simplifying notational assumption, the transition model has

$$\mu_{it} = F(\mathbf{x}_{it}\beta_t) \quad \text{and} \quad \kappa_{it} = F(\mathbf{x}_{it}\beta_t^*), \tag{16}$$

where $F(\cdot)$ – in this paper – is a logistic function, linking the probabilities and the covariates. If we substitute these logit models into equation (15), we obtain a Markov transition model that can be used to estimate covariate-dependent probabilities that vary across subjects and across time. Note that the regression parameters in equation (16) are allowed to vary over time, which in some applications may lead to the presence of unidentifiable parameters. A restriction is then needed to secure identification, for example, the time-stationary assumption $\beta_t = \beta$ and $\beta_t^* = \beta^*$ for all t.

The standard model equation (15) can be extended in several ways to deal with the initial conditions problem, covariates with unknown past history, and unobserved heterogeneity. When using repeated cross-sections, we typically start observing subjects when the process in question is already in progress. Consequently, the first observation will depend on the dependent variable in the period before the sample starts. If we index equation (15) by i, p_{i1} refers to the first observed outcome and p_{i0} to the value of the state prior to the first outcome. While it is difficult to incorporate the prior state, we may "solve" the initial conditions problem by assuming a flexible functional form for the distribution of the first observation. That is, we use a separate logistic function, $P(Y_{i1} = 1) = F(\mathbf{x}_{it}\delta)$, for the cross-section at $t = 1$. In applications it may happen that some covariates are not observable in the past by backcasting. The model may then be modified to include two different sets of parameters for both μ_{it} and κ_{it}, one for the current transition probability estimates and the other for the preceding estimates. We denote by $\mathbf{v}_{it}$ the time-dependent covariate having unknown past history, with parameter vector ζ representing the effect on μ_{it}. We then have, assuming stationary parameters, $\text{logit}(\mu_{it}) = \mathbf{x}_{it}\beta$ for the cross-sections 1, ..., $t-1$ and $\text{logit}(\mu_{it}) = \mathbf{x}_{it}\beta^{**} + \mathbf{v}_{it}\zeta$ for cross-section t, similarly for κ_{it}. This specification allows one to express the current transition probability estimates as a logistic function of both backcastable and nonbackcastable covariates. Further, we may attempt to account for time-invariant omitted variables (i.e., unobserved heterogeneity) by including a unit-specific random error term ε_i in the linear predictor of μ_{it} and κ_{it}. In this logistic-mixture model, we have $\text{logit}(\mu_{it}) = \mathbf{x}_{it}\beta + \gamma_0\varepsilon_i$ and $\text{logit}(\kappa_{it}) = \mathbf{x}_{it}\beta^* + \gamma_1\varepsilon_i$, where γ_0 and γ_1 are the coefficients of the random variable ε_i with zero mean and unit variance. The Gauss-Hermite quadrature approach can be used to numerically integrate the marginal likelihood with respect to the distribution of ε_i. For estimation details, see Pelzer, Eisinga, and Franses (2002). Finally, we note that the model can include unit-specific covariates that, in principle, could have different values for all sampled subjects. In this sense, the model has the potential to account for individual-specific sources of heterogeneity that might be "averaged out" when using aggregate data. However, if all measured covariates are discrete, as in the empirical illustration below, subjects may be segmented into groups comprising individuals with identical observed characteristics. Model construction and parameter estimation with grouped frequency data proceed in the same way as in the individual-level data context.

ML Estimation

Assume that the data consist of T independent cross-sectional drawings. Each cross-section t has m_t different groups of individual units with a unique pattern of values on a set of L covariates and each group i has n_{it} individual units, with $n_{it} \geq 1$. For each individual unit, $Y_{it} \sim Bern\,(p_{it})$ with p_{it} defined by equation (15). Under this sampling scheme, the log likelihood function is

$$\ell\ell = \sum_{t=1}^{T}\sum_{i=1}^{m_t}\left\{\ln\binom{n_{it}}{y_{it}} + y_{it}\ln p_{it} + (n_{it}-y_{it})\ln(1-p_{it})\right\}. \quad (17)$$

The maximum likelihood estimators (MLEs) of the parameters, denoted by the vector θ say, are obtained by maximizing the probability (likelihood) of the sample data, using the derivatives of equation (17) with respect to θ. Obtaining the log likelihood derivatives is conceptually straightforward, but tedious. Let $\eta_{it} = \kappa_{it} - \mu_{it}$. If we omit index i for notational convenience, the first-order derivatives can be expressed as

$$\partial\ell\ell_t/\partial\theta = (y_t - n_t p_t)/p_t(1-p_t)\cdot\partial p_t/\partial\theta, \quad (18)$$

where $\partial p_t/\partial\theta = \partial\mu_t/\partial\theta + (\partial p_{t-1}/\partial\theta)\eta_t + p_{t-1}(\partial\eta_t/\partial\theta)$. If θ is used to estimate μ_t, then $\partial\mu_t/\partial\theta = \mathbf{x}_t\mu_t(1-\mu_t)$ and $\partial\eta_t/\partial\theta = -\partial\mu_t/\partial\theta$. If θ is used to estimate κ_t, then $\partial\mu_t/\partial\theta = \mathbf{0}$ and $\partial\eta_t/\partial\theta = \mathbf{x}_t\kappa_t(1-\kappa_t)$. The derivative values for $\partial p_t/\partial\theta$ can be obtained by recursive substitution, setting $p_0 = 0$ and $\partial p_0/\partial\theta = \mathbf{0}$, and starting with $\partial p_1/\partial\theta = \partial\mu_1/\partial\theta = \mathbf{x}_1\mu_1(1-\mu_1)$. The second-order derivatives are given by

$$\frac{\partial^2\ell\ell_t}{\partial\theta\partial\theta'} = -\frac{(y_t - n_t p_t)^2}{p_t^2(1-p_t)^2}\cdot\frac{\partial p_t}{\partial\theta}\cdot\frac{\partial p_t}{\partial\theta'} + \frac{y_t - n_t p_t}{p_t(1-p_t)}\cdot\frac{\partial^2 p_t}{\partial\theta\partial\theta'}, \quad (19)$$

where

$$\frac{\partial^2 p_t}{\partial\theta\partial\theta'} = \frac{\partial^2 p_{t-1}}{\partial\theta\partial\theta'}\cdot\eta_t + \frac{\partial p_{t-1}}{\partial\theta'}\cdot\frac{\partial\eta_t}{\partial\theta} + \frac{\partial^2\mu_t}{\partial\theta\partial\theta'}\cdot(1-p_{t-1}) - \frac{\partial\mu_t}{\partial\theta'}\cdot\frac{\partial p_{t-1}}{\partial\theta} \quad (20)$$

and $\partial^2\mu_t/\partial\theta\partial\theta' = \mathbf{x}_t'\mathbf{x}_t\mu_t(1-\mu_t)(1-2\mu_t)$. Again, if we set $\partial^2 p_0/\partial\theta\partial\theta' = \partial p_0/\partial\theta = \partial p_0/\partial\theta' = \mathbf{0}$, the values for $\partial^2 p_t/\partial\theta\partial\theta'$ can be obtained recursively, starting with $\partial^2 p_1/\partial\theta\partial\theta' = \partial^2\mu_1/\partial\theta\partial\theta'$. The MLEs can be estimated in the usual way by Fisher's method-of-scoring or Newton's method using the expected or the observed Hessian matrix, respectively. This paper uses the expected Hessian matrix for estimation and inference. An estimated asymptotic variance-covariance matrix is given by the inverse Fisher information matrix for θ evaluated at the ML parameter estimates.

An important justification for the use of ML estimation is that asymptotically ML estimators are unbiased, consistent, and efficient. Unfortunately, the size of the sample necessary to closely approximate these asymptotic properties is essentially unknown for the model in question. We, therefore, also examine the sampling distribution of the parameters using three simulation methods: parametric bootstrap, nonparametric bootstrap, and Bayesian simulation. The parametric bootstrap method uses the observed covariates and the parameters of the fitted model to generate bootstrap replications of the original RCS data (Davison & Hinkley, 1997). The nonparametric bootstrap uses repeated samples from the RCS data. This is done by sampling with replacement. Bayesian simulation is implemented by Markov chain Monte Carlo (MCMC) sampling using the Metropolis algorithm. The Metropolis sampler generates samples from the multivariate posterior distribution of the parameters given the data and a prior density, which is taken to be noninformative in the application given below. For algorithm details, see Pelzer et al. (2002, 2003). The results of these simulations are used to empirically approximate the probability distribution of the parameters and to examine their standard deviation (i.e., estimated standard error) and other distributional properties (e.g., bias, normality, confidence, and credibility intervals).

Application

Data

We illustrate the Markov model with annual data from a German three-wave panel survey on secondary school pupils' interest in learning physics, made available by Vermunt, Langeheine, and Böckenholt (1999). The data were assembled to study the effects of gender (0 = boys, 1 = girls) and school grade in physics (0 = low, 1 = high) on the response variable interest in learning physics (0 = low, 1 = high) at school. The sample available for analysis consists of 541 pupils with complete measurements on all variables in all waves. It is important to note that the three-wave panel data are treated here as if they were three independent cross-sectional samples. The complete data analyzed in this paper are tabulated in Table 1.

As can been seen, there is no information available in the data used for analysis that allows lagged values of y_{it} to be identified. Gender is a time-constant and physics grade a time-varying covariate. In the analysis below, only current values of the covariate school grade in physics are used as predictors. Further, the discrete variables gender and current grade in physics are combined into composite indicators, without the loss of information about these variables. The result is a set of 12 predictor variables (four for each time period) and a total of 20 (= 4 + 2 × 4 + 2 × 4) parameters to be estimated in a nonstationary full model. Note that this is a main-effect-only model and not a saturated one. A saturated model also includes lagged values of physics grade as predictors. The indicator variables are labeled by the letters b (boys) or g (girls) followed by l (low grades) or h (high grades) and a time index t. For example, gh_1 refers to girls with a high grade in physics at $t = 1$. Using these labels, the nonstationary full model can be represented as

Table 1. Pupils' interest in physics data (n of cases = 541)

Gender		0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Physics grade at time t	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
	2	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
	3	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
n		12	15	9	14	18	20	35	151	39	15	12	26	22	15	27	111
y_t	1	3	3	5	8	9	13	22	100	6	1	2	8	5	7	9	48
	2	1	7	5	10	7	7	19	113	3	0	0	7	1	3	9	48
	3	3	9	5	10	4	7	11	103	4	0	2	8	2	5	7	40

Note. Gender (0 = boys, 1 = girls), physics grade (0 = low, 1 = high) at time t = 1, 2, 3, n is the number of pupils, and y_t the number of pupils with high interest in physics at t. Reprinted in part from Vermunt et al. (1999) with permission.

$$\begin{aligned}
\text{logit}(p_1) &= \delta_1 + \delta_2 \text{bl}_1 + \delta_3 \text{gl}_1 + \delta_4 \text{gh}_1 \\
\text{logit}(\mu_2) &= \beta_1 \text{bl}_2 + \beta_2 \text{gl}_2 + \beta_3 \text{bh}_2 + \beta_4 \text{gh}_2 \\
\text{logit}(\kappa_2) &= \beta_1^* \text{bl}_2 + \beta_2^* \text{gl}_2 + \beta_3^* \text{bh}_2 + \beta_4^* \text{gh}_2 \\
\text{logit}(\mu_3) &= \beta_5 \text{bl}_3 + \beta_6 \text{gl}_3 + \beta_7 \text{bh}_3 + \beta_8 \text{gh}_3 \\
\text{logit}(\kappa_3) &= \beta_5^* \text{bl}_3 + \beta_6^* \text{gl}_3 + \beta_7^* \text{bh}_3 + \beta_8^* \text{gh}_3.
\end{aligned}$$

The constant term for $\text{logit}(p_1)$ designates the (omitted) category of boys with a high grade at $t = 1$. The reason for omitting this particular category will become evident below. The other four equations have no constant term. The reason for omitting them is that it admits an interpretation of the parameters in terms of logits (i.e., log odds-ratios).

Model Selection and Parameter Estimation

The analysis consists of determining a parsimonious transition model that fits the observed data well. Prior to this model selection process, two pre-analyses were performed. One was to determine a parsimonious representation of the data for cross-section $t = 1$. The following model appears best fitting and the most parsimonious parameterization to estimate p_1: $\text{logit}(p_1) = \delta_1 + \delta_2 \text{bl}_1 + 2\,\delta_2 \text{gl}_1 + \delta_2 \text{gh}_1$. The model includes two parameters (δ_1 and δ_2) and it selects (the most populated category of) boys with a high grade at $t = 1$ as the reference category. The other data exploration concerned the lower and upper bounds for μ_t and κ_t for each gender-physics grade category at $t = 2$ and $t = 3$. These (nondeterministic) bounds provide meaningful support in the model selection process. To compute the bounds, we assume that the unobserved marginal proportions at $t - 1$ for cross-section t are equal to the observed marginal proportions of the corresponding cases in cross-section $t - 1$, where corresponding means the same (lagged) covariate values at $t - 1$. For time-constant covariates, f in equations (12) and (13) is taken to be $(y_{t-1}/n_{t-1}) \times n_t$, where n_t and n_{t-1} are the number of cases at t and $t - 1$, respectively, and y_{t-1}, the observed counts for cases with identical covariate values at $t - 1$. In this application, the covariates are time-varying however, and pupils with a particular covariate value at time point t have either one of two covariate values at $t - 1$. We, therefore, take f to be the weighted sum of the two observed marginal proportions at $t - 1$, denoted by (y^0_{t-1}/n^0_{t-1}) and (y^1_{t-1}/n^1_{t-1}), where the weights are the number of cases at t with identical covariate values at $t - 1$, denoted by n^0_{t-1} and n^1_{t-1}. Thus, $f = (y^0_{t-1}/n^0_{t-1}) \times n^0_t + (y^1_{t-1}/n^1_{t-1}) \times n^1_t$. For example, for boys with low grades at $t = 2$ (and either low or high grades at $t = 1$) we have $f = (19/50) \times 27 + (144/224) \times 38 = 35$. Taken the observed and imputed margins as given, bounds for μ and κ at $t = 2$ and $t = 3$ were obtained for each gender-physics grade category. As indicated, we cannot be entirely confident that the true lower and upper limits are being revealed by the sample. We, therefore, applied the bootstrap and generated 5,000 samples (by sampling with replacement) from the observed margins. The resulting (observed and imputed) marginal counts were then employed to construct bootstrap confidence intervals for the lower and upper bounds using

Table 2. Lower and upper bounds for μ_t and κ_t

	μ_2		μ_3		κ_2		κ_3	
	L	U	L	U	L	U	L	U
Boys – low grade	0	0.73	0	0.70	0	0.63	0	0.56
	[0	1]	[0	0.97]	[0	0.88]	[0	0.77]
Girls – low grade	0	0.11	0	0.18	0	0.28	0	0.79
	[0	0.19]	[0	0.28]	[0	0.53]	[0	1]
Boys – high grade	0.23	1	0.02	1	0.52	1	0.44	1
	[0	1]	[0	1]	[0.40	1]	[0.33	1]
Girls – high grade	0.02	0.56	0.01	0.46	0	1	0	1
	[0	0.69]	[0	0.58]	[0	1]	[0	1]

Note. 95% bootstrap confidence intervals in brackets.

the percentile method. Table 2 reports the bounds observed in the sample along with the endpoints of the 95% bootstrap confidence intervals.

As can be seen, for a majority of the categories the intervals are narrower than [0, 1]. Some of the intervals are simply too wide to provide useful information, but others are narrow enough to provide meaningful support in model selection. Note that for girls with a low grade in physics at $t = 2$, samples' lower and upper bounds for μ_2 are [0, 0.11], implying a low 0–1 transition probability at $t = 2$. Also note that for girls with low grades the interval for κ becomes substantially wider over time, even if we take the bootstrap findings into account. This hints at the possibility that the parameters for this category are time-varying.

After pre-processing, the data were submitted to a series of analyses to select a parsimonious transition model. The selection was carried out in a step-by-step fashion, beginning with a time-stationary model with time-invariant parameters for categories with the same gender-grade value ($\beta_1 = \beta_5$, $\beta_2 = \beta_6$, etc.). We subsequently extended the model to allow for time-varying coefficients by examining the parameters one-by-one. The resulting nonstationary model was then progressively simplified by fixing parameters to zero or constraining them to be equal. The nested model with fewer parameters was accepted if the increase in deviance (i.e., $-2\Delta\ell\ell$) was not statistically significant at the 0.10 level. The results of this modeling process are presented in Table 3.

Table 3. Model determination by forward selection and backward elimination

				p_1		μ_t				κ_t			
		$-\ell\ell$	#	δ_1	δ_2	β_1	β_2	β_3	β_4	β_1^*	β_2^*	β_3^*	β_4^*
Model						β_5	β_6	β_7	β_8	β_5^*	β_6^*	β_7^*	β_8^*
1	Time stationary model	963.64	10	0.63	−1.07	−1.36	−3.41	−0.37	−2.77	−0.37	−0.40	1.63	2.29
						−1.36	−3.41	−0.37	−2.77	−0.37	−0.40	1.63	2.29
2	$\beta_1 \neq \beta_5$	963.63	11	0.63	−1.07	−1.26	−3.41	−0.38	−2.77	−0.38	−0.40	1.64	2.29
						−1.42	—	—	—	—	—	—	—
3	$\beta_2 \neq \beta_6$	961.86	11	0.63	−1.06	−1.36	−21.36	−0.38	−2.66	−0.37	−0.60	1.63	2.12
						—	−2.30	—	—	—	—	—	—
4	$\beta_3 \neq \beta_7$	962.50	11	0.60	−1.05	−1.38	−3.43	0.18	−2.78	−0.34	−0.39	1.47	2.38
						—	—	−0.68	—	—	—	—	—
5	$\beta_4 \neq \beta_8$	963.36	11	0.63	−1.09	−1.35	−3.58	−0.36	−2.20	−0.37	−0.29	1.62	2.16
						—	—	—	−3.69	—	—	—	—
6	$\beta_1^* \neq \beta_5^*$	963.64	11	0.63	−1.07	−1.36	−3.41	−0.37	−2.77	−0.39	−0.40	1.63	2.29
						—	—	—	—	−0.35	—	—	—
7	$\beta_2^* \neq \beta_6^*$	961.87	11	0.63	−1.05	−1.36	−3.42	−0.38	−2.50	−0.37	−1.10	1.64	1.87
						—	—	—	—	—	0.61	—	—
8	$\beta_3^* \neq \beta_7^*$	963.05	11	0.60	−1.05	−1.35	−3.43	−0.40	−2.82	−0.37	−0.39	2.07	2.37
						—	—	—	—	—	—	1.34	—
9	$\beta_4^* \neq \beta_8^*$	963.39	11	0.62	−1.08	−1.35	−3.42	−0.36	−2.66	−0.37	−0.39	1.62	3.33
						—	—	—	—	—	—	—	1.49
10	$\beta_2 \neq \beta_6$	960.31	13	0.60	−1.04	−1.39	−20.06	0.18	−2.58	−0.33	−0.80	1.47	2.02
	$\beta_3 \neq \beta_7$, $\beta_2^* \neq \beta_6^*$					—	−2.69	−0.68	—	—	0.17	—	—
11	$\beta_2 = 5\beta_6$	960.31	12	0.60	−1.04	−1.39	*	0.18	−2.58	−0.33	−0.80	1.47	2.02
						—	−2.69	−0.68	—	—	0.17	—	—
12	$\beta_3 = 0$	960.34	11	0.60	−1.04	−1.34	*	*0*	−2.58	−0.35	−0.80	1.60	2.01
						—	−2.69	−0.78	—	—	0.17	—	—
13	$\beta_6^* = 0$	960.37	10	0.60	−1.04	−1.34	*	*0*	−2.58	−0.35	−0.80	1.60	2.02
						—	−2.59	−0.78	—	—	*0*	—	—
14	$\beta_6 = \beta_4$	960.37	9	0.60	−1.04	−1.34	*	*0*	−2.59	−0.35	−0.80	1.60	2.03
						—	−2.59	−0.78	—	—	*0*	—	—
15	$\beta_7 = \beta_1$	960.52	8	0.60	−1.03	−1.01	*	*0*	−2.59	−0.51	−0.80	1.67	2.04
						—	−2.59	−1.01	—	—	*0*	—	—
16	$\beta_2^* = \beta_1^*$	960.63	7	0.60	−1.03	−0.93	*	*0*	−2.65	−0.62	−0.62	1.64	2.12
						—	−2.65	−0.93	—	—	*0*	—	—
17	$\beta_4^* = \beta_3^*$	960.82	6	0.60	−1.04	−0.94	*	*0*	−2.44	−0.64	−0.64	1.69	1.69
						—	−2.44	−0.94	—	—	*0*	—	—

Note. # is the number of parameters. Long dashes (—) denote parameter values identical to the estimate given immediately above. Asterisks (*) for β_2 denote parameter values five times to the estimate given immediately below. Zeros in italic (*0*) indicate parameters fixed to 0 (implying a probability of 0.5).

The table reports the minus log likelihood values ($-\ell\ell$), the number of parameters (#), the estimated effects on the marginal probability at $t = 1$ (δ_1 and δ_2), the transition probability from low to high interest in physics (β_1–β_8), and the probability to stay in the high interest state (β_1^*–β_8^*). The top row refers to the time stationary model (model 1). The figures for p_1 indicate that boys with high grades at $t = 1$ are more likely to show high interest in physics than both boys with low grades and girls. Girls with low grades at $t = 1$ are least likely to be highly interested in learning physics (recall that the parameter pertaining to this category is $2\delta_2$). The results for μ_t imply that boys with high grades in physics are most likely to change from low to high interest. The estimated logits are the smallest for girls with low grades. The parameter estimates with respect to κ_t reveal that the probability of 1–1 transition is high for pupils with a high grade in physics and relatively low for pupils with a low grade. It is also useful to note that the time-stationary transition model 1, which yielded $\ell\ell = -963.64$, is superior to a marginal model of the data, which has a log likelihood value of -971.88.

Models 2–9 test for nonstationarity of the parameters one-by-one. The test results show that removing the restriction of time-invariance does not significantly improve the fit of the model compared to model 1, except for three equality constraints (models 3, 4, and 7). The assumptions $\beta_2 = \beta_6$, $\beta_3 = \beta_7$, and $\beta_2^* = \beta_6^*$ significantly deteriorate the fit, although not substantially. These findings imply that both μ and κ are constant over time for all groups, except for μ of boys with high grades and μ and κ of girls with low grades. The latter comes as no surprise given the lower and upper bounds for this category.

Models 10–13 further examine the three sets of time-varying parameters. Model 10 removes the restrictions simultaneously, with equivalent results. The findings show that the estimate for β_2 has a (arbitrary) large negative value, implying (near) zero probability. In model 11, β_2 is set equal

Table 4. Estimated expected total (top) and conditional (bottom) probabilities

	μ_2^*				μ_3^*				κ_2^*				κ_3^*			
Model	bl	gl	bh	gh	bl	gl	bh	gh	bl	gl	bh	gh	bl	gl	bh	gh
1	0.09	0.02	0.15	0.04	0.10	0.03	0.16	0.04	0.22	0.11	0.52	0.31	0.21	0.08	0.51	0.29
2	0.10	0.02	0.15	0.04	0.09	0.03	0.16	0.04	0.22	0.11	0.52	0.31	0.21	0.08	0.51	0.29
3	0.09	0.00	0.15	0.04	0.10	0.07	0.16	0.04	0.22	0.10	0.52	0.31	0.21	0.07	0.51	0.28
4	0.09	0.02	0.21	0.04	0.09	0.02	0.12	0.04	0.22	0.11	0.50	0.31	0.23	0.08	0.52	0.29
5	0.09	0.02	0.16	0.07	0.10	0.02	0.16	0.02	0.22	0.11	0.52	0.31	0.21	0.09	0.51	0.30
6	0.09	0.02	0.15	0.04	0.10	0.03	0.16	0.04	0.22	0.11	0.52	0.31	0.22	0.08	0.51	0.29
7	0.09	0.02	0.15	0.05	0.10	0.03	0.16	0.05	0.22	0.07	0.52	0.30	0.21	0.12	0.51	0.27
8	0.09	0.02	0.15	0.04	0.09	0.02	0.15	0.04	0.22	0.11	0.55	0.32	0.22	0.08	0.50	0.29
9	0.09	0.02	0.16	0.04	0.10	0.02	0.16	0.04	0.22	0.11	0.52	0.33	0.21	0.09	0.51	0.27
10	0.09	0.00	0.21	0.05	0.09	0.05	0.12	0.05	0.23	0.08	0.50	0.31	0.23	0.10	0.52	0.27
11	0.09	0.00	0.21	0.05	0.09	0.05	0.12	0.05	0.23	0.08	0.50	0.31	0.23	0.10	0.52	0.27
12	0.10	0.00	0.19	0.05	0.09	0.05	0.11	0.05	0.22	0.08	0.51	0.31	0.22	0.10	0.53	0.27
13	0.10	0.00	0.19	0.05	0.09	0.06	0.11	0.05	0.22	0.08	0.51	0.31	0.22	0.09	0.53	0.27
14	0.10	0.00	0.19	0.05	0.09	0.06	0.11	0.05	0.22	0.08	0.51	0.31	0.22	0.09	0.53	0.27
15	0.12	0.00	0.19	0.05	0.12	0.06	0.09	0.05	0.20	0.08	0.52	0.31	0.21	0.09	0.54	0.27
16	0.13	0.00	0.19	0.04	0.13	0.05	0.10	0.05	0.19	0.09	0.52	0.31	0.19	0.09	0.54	0.28
17	0.13	0.00	0.19	0.05	0.13	0.07	0.10	0.06	0.19	0.09	0.52	0.29	0.19	0.09	0.54	0.26
	μ_2				μ_3				κ_2				κ_3			
1	0.20	0.03	0.41	0.06	0.20	0.03	0.41	0.06	0.41	0.40	0.84	0.91	0.41	0.40	0.84	0.91
2	0.22	0.03	0.41	0.06	0.19	0.03	0.41	0.06	0.41	0.40	0.84	0.91	0.41	0.40	0.84	0.91
3	0.20	0.00	0.41	0.07	0.20	0.09	0.41	0.07	0.41	0.35	0.84	0.89	0.41	0.35	0.84	0.89
4	0.20	0.03	0.55	0.06	0.20	0.03	0.34	0.06	0.42	0.40	0.81	0.91	0.42	0.40	0.81	0.91
5	0.21	0.03	0.41	0.10	0.21	0.03	0.41	0.02	0.41	0.43	0.83	0.90	0.41	0.43	0.83	0.90
6	0.20	0.03	0.41	0.06	0.20	0.03	0.41	0.06	0.40	0.40	0.84	0.91	0.41	0.40	0.84	0.91
7	0.20	0.03	0.41	0.08	0.20	0.03	0.41	0.08	0.41	0.25	0.84	0.87	0.41	0.65	0.84	0.87
8	0.21	0.03	0.40	0.06	0.21	0.03	0.40	0.06	0.41	0.40	0.89	0.91	0.41	0.40	0.79	0.91
9	0.21	0.03	0.41	0.07	0.21	0.03	0.41	0.07	0.41	0.40	0.84	0.97	0.41	0.40	0.84	0.82
10	0.20	0.00	0.54	0.07	0.20	0.06	0.34	0.07	0.42	0.31	0.81	0.88	0.42	0.54	0.81	0.88
11	0.20	0.00	0.54	0.07	0.20	0.06	0.34	0.07	0.42	0.31	0.81	0.88	0.42	0.54	0.81	0.88
12	0.21	0.00	0.50	0.07	0.21	0.06	0.31	0.07	0.41	0.31	0.83	0.88	0.41	0.54	0.83	0.88
13	0.21	0.00	0.50	0.07	0.21	0.07	0.31	0.07	0.41	0.31	0.83	0.88	0.41	0.50	0.83	0.88
14	0.21	0.00	0.50	0.07	0.21	0.07	0.31	0.07	0.41	0.31	0.83	0.88	0.41	0.50	0.83	0.88
15	0.27	0.00	0.50	0.07	0.27	0.07	0.27	0.07	0.38	0.31	0.84	0.88	0.38	0.50	0.84	0.88
16	0.28	0.00	0.50	0.07	0.28	0.07	0.28	0.07	0.35	0.35	0.84	0.89	0.35	0.50	0.84	0.89
17	0.28	0.00	0.50	0.08	0.28	0.08	0.28	0.08	0.35	0.35	0.84	0.84	0.35	0.50	0.84	0.84

Note. bl = boys with low grades, gl = girls with low grades, bh = boys with high grades, gh = girls with high grades; $\mu_t^* = n_t^{-1} \sum (1 - p_{t-1})\, \mu_t$ and $\kappa_t^* = n_t^{-1} \sum p_{t-1}\, \kappa_t$.

to $5\beta_6$, without any loss in log likelihood. Both model 12 and model 13 constrain one element of the pair of time-varying parameters to zero. The findings reveal that the restricted models fit the data equally well.

Models 14–17 additionally reduce the number of estimated parameters using equality constraints. For example, model 14 tests the assumption that μ_3 for girls with a low grade equals $\mu_2 = \mu_3$ for girls with a high grade (i.e., $\beta_6 = \beta_4$). Looking at the log likelihood changes that result from removing parameters from the model, it can be seen that none of the increments in $\ell\ell$ is significant. The finally selected model (model 17) has only six parameters to be estimated. Elimination of any of the remaining model parameters would result in a significant increase in deviance, except for one equality constraint, namely $\beta_1^* = -\beta_1$. This constraint was not imposed because it concerns a cross-restriction between elements of two different parameter sets (i.e., μ_t and κ_t). Also, the presence of this restriction has hardly any effect on the results reported here. The inclusion of two random residual effect parameters, γ_0 and γ_1, to account for lack of fit due to extra binomial variation does not significantly change the results either. The difference in deviance between the (selected) model that includes and the model that excludes the ancillary variance parameters is $-2\Delta\ell\ell = 0.441$. Although traditional likelihood-ratio testing should not be used to test the significance of the random effects – as it cannot be assumed to have a χ^2 distribution – the improvement in fit is not worth the increase in model complexity.

Also note, from Table 3, that adding parameters to the model (models 1–9) has almost no effect and eliminating them (models 10–17) has only small effect on the other parameter estimates. This can also be seen by examining the estimated expected probabilities given in Table 4.

Data in Table 4 monitors the total and conditional estimated expected probabilities of all models. The total probabilities are given by $\mu_t^* = n_t^{-1} \sum (1 - p_{t-1})\, \mu_t$ and $\kappa_t^* = n_t^{-1} \sum p_{t-1}\, \kappa_t$. Simplifying the model produces no substantial changes in the predicted transition probabilities, either in the total or in the conditional probabilities, except perhaps for κ_2 and κ_3 for girls with a low grade in physics.

Bootstrap and MCMC Simulation

The ML parameter estimates of the selected model (model 17) and the standard deviations are presented in the second column of Table 5. To empirically approximate the probability distribution of the parameter estimates 5,000 parametric bootstrap samples were obtained from the data using the selected parametric model. These replicate data sets were subsequently analyzed, using ML. The third column of Table 5 provides the mean and the sample standard deviation of the parametric bootstrap estimates for each parameter.

As shown in Table 5, the mean values are close to the Markov parameter estimates for the RCS data and the sample standard deviations are similar to the Fisher information-based standard errors. The RCS Markov estimates appear to be only slightly biased, with the largest absolute value being 0.055 for β_4. A frequently applied rule of thumb is that a good estimator should be biased by less than 25% of its standard deviation (Efron & Tibshirani, 1993). The ratios of estimated bias to standard deviation are all (much) less than 0.25. The fourth column of the table presents the nonparametric bootstrap findings obtained by drawing (with replacement) 5,000 samples from the data. One can note from this table that the means and standard deviations of the parameter estimates are roughly similar to both the parametric bootstrap and the RCS Markov results. The MCMC findings were obtained using the Metropolis sampler, wherein the regression parameters were assumed to follow a multivariate normal distribution with diffuse or noninformative priors. In the Markov chain sampling, we ran the algorithm 1,000,000 times excluding an initial burn-in of 4,000 samples. Several diagnostics were used to assess convergence to the posterior, using a thinning interval of 10 (Geweke *Z*-test, Heidelberg and Welsh stationarity and interval halfwidth tests, Raftery and Lewis convergence diagnostic). None of these diagnostics indicates convergence failure for any of the sampled parameters. The posterior estimates are shown in Table 5. Inspection of the mean of the 10^6 samples reveals that there are no gross discrepancies in magnitude compared to the ML estimates for the original sample. The MCMC standard deviations and the RCS Markov standard errors are also similar to one another. Thus, both Bayesian and frequentist methods for obtaining estimates seem to indicate that the ML point estimates of the parameters are rather accurate in this application and that the inverse of the Fisher information matrix may be used as a good estimate of the covariance matrix of the parameter estimates.

Table 5. ML and MCMC estimates (n of obs = 1,623)

	RCS Markov	Parametric bootstrap	Nonparametric bootstrap	MCMC	Cross-section subsamples
p_1					
δ_1	0.599	0.604	0.603	0.593	0.604
	(0.127)	(0.128)	(0.128)	(0.127)	(0.105)
δ_2	−1.036	−1.042	−1.043	−1.029	−1.047
	(0.126)	(0.127)	(0.128)	(0.127)	(0.102)
μ_t					
β_1	−0.941	−0.973	−0.970	−0.991	−0.973
	(0.368)	(0.391)	(0.399)	(0.393)	(0.316)
β_4	−2.437	−2.492	−2.498	−2.539	−2.523
	(0.355)	(0.397)	(0.419)	(0.564)	(0.345)
κ_t					
β_1^*	−0.636	−0.655	−0.658	−0.619	−0.693
	(0.319)	(0.330)	(0.316)	(0.318)	(0.264)
β_3^*	1.689	1.702	1.704	1.725	1.711
	(0.242)	(0.245)	(0.248)	(0.252)	(0.204)

Note. The standard error (RCS Markov) and standard deviation (others) are reported in parentheses. The parametric and nonparametric bootstrap findings are based on 5,000 samples from the data and the MCMC findings on 10^6 Metropolis sampler posterior estimates. The mean of the parameter estimates is reported as the point estimate. The rightmost column summarizes the results of the ML analyses of 5,000 "cross-sectional" subsamples from the data. The figure in parentheses in this column is the standard deviation divided by $\sqrt{1,623/540}$.

Independent Cross-Sectional Subsamples

It is also informative to report how the parameter estimates compare to the estimates we get from the independent samples. As indicated, the original data were treated here as a sequence of independent cross-sections. Nevertheless, to eliminate the possibility that the findings are artificially induced by the panel nature of the data, we randomly drew (without replacement) samples of 540 different pupils from the 1,623 panel observations, where each sample consists of three sets – one for each time period – of 180 pupils, and where each pupil is selected only once per sample. The total number of possible "cross-sectional" drawings in this application is on the order of 10^{255}. Since this is beyond any current ability to analyze, we randomly drew 5,000 samples and analyzed each data set separately using ML. The rightmost column of Table 5 reports the average value of the parameters across the samples along with the standard deviations divided by $\sqrt{1,623/540}$. We note that the mean values of the parameters are close to the ML estimates obtained for the original data and that the standard deviations agree well with the estimated standard errors. The results thus demonstrate that the RCS Markov findings are not due to the panel character of the data.

Expected Probabilities and Observed Proportions

To understand how well the model recovers the transitions in the panel, we compare the estimated expected transition probabilities with the observed transitions in the panel. The top part of Table 6 provides the total probabilities and the observed proportions for each gender-physics grade category.

As can be seen, the model recovers the marginal probabilities at $t = 1$ and the total transition probabilities at both $t = 2$ and $t = 3$ quite precisely. The bottom part of Table 6 reports the estimated expected conditional probabilities and the observations in the panel for each of the four categories of gender and current physics grade. The results indicate that the model reproduces most of the prominent features in the data well, despite the small-sized categories. For many categories, the conditional probabilities predicted by the model match with the observations in the panel. An exception is the probability to stay in the high interest state for some categories. The model underestimates $\kappa_2 = \kappa_3$ for boys with low grades in physics and it overestimates $\kappa_2 = \kappa_3$ for girls with high grades. However, the findings overall illustrate that in this application at least the model is well able to recover the panel data observations.

Discussion

This paper is not intended to suggest that a series of one-shot surveys is as informative about longitudinal changes as repeated measurements. Quite obviously, it is not. When we are given the opportunity to choose, individual panel observations are surely to be preferred to repeated cross-sections for analyzing 0–1 transitions over time. The model elaborated here is useful for the situation where we find ourselves unable to obtain serial y_{it-1} observations, and cross-sectional drawings are the only available data source.

Table 6. Estimated expected total (top) and conditional (bottom) probabilities and observed panel proportions

	bl	gl	bh	gh	tot	bl	gl	bh	gh	tot
			p_1							
exp	0.39	0.19	0.65	0.39	0.46					
obs	0.38	0.19	0.64	0.39	0.46					
n	50	92	224	175	541					
			μ_2^*					μ_3^*		
exp	0.13	0.00	0.19	0.05	0.10	0.13	0.07	0.10	0.06	0.08
obs	0.11	0.01	0.18	0.14	0.13	0.04	0.06	0.10	0.10	0.09
n	65	91	209	176	541	74	100	200	167	541
			κ_2^*					κ_3^*		
exp	0.19	0.09	0.52	0.29	0.34	0.19	0.09	0.54	0.26	0.33
obs	0.23	0.07	0.52	0.22	0.31	0.27	0.09	0.55	0.22	0.32
n	65	91	209	176	541	74	100	200	167	541
			μ_2					μ_3		
exp	0.28	0.00	0.50	0.08	0.24	0.28	0.08	0.28	0.08	0.18
obs	0.19	0.01	0.51	0.23	0.24	0.07	0.07	0.32	0.16	0.15
n	37	72	74	109	292	42	87	63	109	301
			κ_2					κ_3		
exp	0.35	0.35	0.84	0.84	0.71	0.35	0.50	0.84	0.84	0.73
obs	0.54	0.32	0.81	0.58	0.68	0.63	0.69	0.80	0.62	0.73
n	28	19	135	67	249	32	13	137	58	240

Note. bl = boys with low grades, gl = girls with low grades, bh = boys with high grades, gh = girls with high grades, tot = total; *exp* denotes the estimated expected probabilities, *obs* the observed proportions in the panel data, *n* is the number of cases; $\mu_t^* = n_t^{-1} \sum (1 - p_{t-1})\, \mu_t$ and $\kappa_t^* = n_t^{-1} \sum p_{t-1}\, \kappa_t$.

Although the results of this paper seem promising, they are not conclusive. Obviously, one can draw no firm conclusions about the general properties of models from one particular application. The small-sized data file to which we have applied the model is not representative of the population of data sets to which it might be applied (for additional applications, see Moffitt, 1993; Pelzer & Eisinga, 2002; Pelzer et al., 2001, 2002, 2003). A valuable alternative to the approach presented here would be to carry out an extensive Monte Carlo simulation study in which the ML estimators are compared to the known (true) values under a variety of conditions. Another avenue of further inquiry is maximum entropy estimation (Golan et al., 1996). When using ML some simplifications or approximations are required to make the analysis tractable. This is accomplished by using a pre-specified likelihood function that is based on assumptions about the underlying data-generating process. Also, a parametric structure is imposed here through the choice of a logistic function for μ and κ, although the underlying distribution is rarely, if ever, known in practice. Further, in some applications the model may have to be simplified to secure parameter identification by reducing the number of unknowns. This may affect the usefulness of the model and lead to unfavorable estimation and inference consequences. The principle of maximum entropy offers an alternative method of statistical inference for which such a priori assumptions are not required. It aims at assigning values to probability distributions using minimal distribution assumptions. In maximum entropy estimation, the role of the data is restricted only to the provision of constraints on the set of allowed probability distributions.

Finally, there has been a tremendous accumulation of data obtained from the repeated surveys over the past several decades, but we rarely find models specially designed to analyze these data. We believe that the potential for analysis inherent in the clustering of repeated cross-section with independent samples has not been fully exploited. It is hoped that the present contribution will serve to partly rectify this situation.

Acknowledgements

This research was supported by a grant from the Netherlands Organisation for Scientific Research (NWO), Division for Social Sciences (# 480-04-009). The data on physics education used in this paper were collected by the Institute for Science Education in Kiel (Germany) and reprinted in part from Vermunt, Langeheine, and Böckenholt (1999). We thank Rolf Langeheine for permission to reproduce the data.

References

Abowd, J. M., & Zellner, A. (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics, 3*, 254–283.

Achen, C. H., & Shively, W. P. (1995). *Cross-level inference*. Chicago, Ill: University of Chicago Press.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science, 7*, 131–153.

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika, 34*, 123–138.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Böckenholt, U., & Dillon, W. R. (2000). Inferring latent brand dependencies. *Journal of Marketing Research, 37*, 72–87.

Chambers, R. L., & Steel, D. G. (2001). Simple methods for ecological inference in 2×2 tables. *Journal of the Royal Statistical Society, Series A, 164*, 175–192.

Cross, P. J., & Manski, C. F. (2002). Regressions, short and long. *Econometrica, 70*, 357–368.

Davison, A. C., & Hinkley, D. V. (1997), *Bootstrap methods and their application*. Cambridge, MA: Cambridge University Press.

Dobra, A., Tebaldi, C., & West, M. (2003). *Bayesian inference in incomplete multi-way tables*. Durham, NC: Institute of Statistics and Decision Sciences, Duke University.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eisinga, R. (2008). Information loss for 2×2 tables with missing cell counts: Binomial case. *Statistica Neerlandica, 62*, 239–254.

Fienberg, S. (1997). *Confidentiality and disclosure limitation methodology. Challenges for national statistics and statistical research*. Pittsburgh, PA: Department of Statistics, Carnegie Mellon University.

Fingleton, B. (1997). Specification and testing of Markov chain models: An application to convergence in the European Union. *Oxford Bulletin of Economics and Statistics, 59*, 385–403.

Fisher, R. A. (1935). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society, 98*, 39–82.

Golan, A., Judge, G., & Miller, D. (1996). *Maximum entropy econometrics. Robust estimation with limited data*. New York: Wiley.

Golan, A., Judge, G., & Robinson, S. (1994). Recovering information from incomplete or partial multisectoral economic data. *Review of Economics and Statistics, 76*, 541–549.

Haber, M. (1989). Do the marginal totals of a contingency table contain information regarding the table proportions? *Communications in Statistics – Theory and Methods, 18*, 147–156.

Hamdan, M. A., & Nasro, M. O. (1986). Maximum likelihood estimation of the parameters of the bivariate binomial distribution. *Communication in Statistics – Theory and Methods, 15*, 747–754.

Hawkins, D. L., Han, C. P., & Eisenfeld, J. (1996). Estimating transition probabilities from aggregate samples augmented by haphazard recaptures. *Biometrics, 52*, 625–638.

Judge, G., Miller, D., & Tam Cho, W. K. (2003). An information theoretic approach to ecological estimation and inference. In G. King, O. Rosen, & M. Tanner (Eds.), *Ecological inference. New methodological strategies* (pp. 162–187). New York: Cambridge University Press.

Kalbfleish, J. D., & Lawless, J. F. (1984). Least squares estimation of transition probabilities from aggregate data. *Canadian Journal of Statistics, 12*, 169–182.

Kalbfleish, J. D., & Lawless, J. F. (1985). The analysis of panel data under a Markovian assumption. *Journal of the American Statistical Association, 80*, 863–871.

Karantininis, K. (2002). Information-based estimators for the non-stationary transition probability matrix: An application to the Danish pork industry. *Journal of Econometrics, 107*, 275–290.

Kelton, C. M. L. (1981). Estimation of time-independent Markov processes with aggregate data: A comparison of techniques. *Econometrica, 49*, 517–518.

Kelton, W. D., & Kelton, C. M. L. (1984). Hypothesis tests for Markov process models estimated from aggregate frequency data. *Journal of the American Statistical Association, 79*, 922–928.

King, G. (1997). *A solution to the ecological inference problem. Reconstructing individual behavior from aggregate data*. Cambridge, MA: Cambridge University Press.

King, G., Rosen, O., & Tanner, M. (2003). *Ecological inference. New methodological strategies*. New York: Cambridge University Press.

Kocherlakota, S., & Kocherlakota, K. (1992), *Bivariate discrete distributions*. New York: Marcel Dekker.

Lawless, J. F., & McLeish, D. (1984). The information in aggregate data from Markov chains. *Biometrika, 71*, 419–430.

Lee, T. C., Judge, G. G., & Zellner, A. (1970). *Estimating the parameters of the Markov probability model from aggregate time series data*. Amsterdam: North-Holland.

Li, W. K., & Kwok, M. C. O. (1990). Some results on the estimation of a higher order Markov chain. *Communications in Statistics. Part B. Simulation and Computation, 19*, 363–380.

MacRae, E. C. (1977). Estimation of time-varying Markov processes with aggregate data. *Econometrica, 45*, 183–198.

McCue, K. F. (1995). *Individual choice and ecological analysis*. Pasadena, CA: California Institute of Technology.

McCullagh, P., & Nelder, J. A. (1992). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Moffitt, R. (1990). The effect of the U.S. welfare system on marital status. *Journal of Public Economics, 41*, 101–124.

Moffitt, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics, 59*, 99–123.

Pelzer, B., & Eisinga, R. (2002). Bayesian estimation of transition probabilities from repeated cross sections. *Statistica Neerlandica, 56*, 23–33.

Pelzer, B., Eisinga, R., & Franses, P. H. (2001). Estimating transition probabilities from a time series of repeated cross sections. *Statistica Neerlandica, 55*, 248–261.

Pelzer, B., Eisinga, R., & Franses, P. H. (2002). Inferring transition probabilities from repeated cross sections. *Political Analysis, 10*, 113–133.

Pelzer, B., Eisinga, R., & Franses, P. H. (2003). Ecological panel inference from repeated cross sections. In G. King, O. Rosen, & M. Tanner (Eds.), *Ecological inference. New methodological strategies* (pp. 188–205). New York: Cambridge University Press.

Plackett, R. L. (1977). The marginal totals of a 2×2 table. *Biometrika, 64*, 37–42.

Richardson, S., & Montfort, C. (2000). Ecological correlation studies. In P. Elliott, J. C. Wakefield, N. G. Best, & D. J. Briggs (Eds.), *Spatial epidemiology. Methods and applications* (pp. 205–220). Oxford: Oxford University Press.

Tebaldi, C., & West, M. (1998). *Reconstruction of contingency tables with missing data*. Durham, NC: Institute of Statistics and Decision Sciences, Duke University.

Vermunt, J. K., Langeheine, R., & Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics, 24*, 179–207.

Wakefield, J. (2003). Prior and likelihood choices in the analysis of ecological data. In G. King, O. Rosen, & M. Tanner (Eds.), *Ecological inference. New methodological strategies* (pp. 13–50). New York: Cambridge University Press.

Woodward, J. A., & Palmer, C. G. S. (1997). On the exact convolution of discrete random variables. *Applied Mathematics and Computation, 83*, 69–77.

Zepeda, L. (1995). Technical change and the structure of production. A nonstationary Markov analysis. *European Review of Agricultural Economics, 22*, 41–60.

Rob Eisinga

Department of Social Science Research Methods
Radboud University Nijmegen
P.O. Box 9104
6500 HE Nijmegen
The Netherlands
Tel. +31 24 3615722
Fax +31 24 3612351
E-mail r.eisinga@maw.ru.nl

